National College of Ireland

MSc Data Analytics

Execution and Examination of Statistical Methodologies

Anand Basant Haritwal
MSC in Data analytics
National College of Ireland
Dublin-1
Student Id- x19174489@student@ncirl.ie

Lecturer

Tony Delaney

28 April 2020

Contents

Abstract

Logistics Regression is used to predict outcomes of categorical variables. In this report total life satisfaction of a person is predicted as satisfied with life or unsatisfied. ANOVA testing is done to compare means of two or more groups. In this report, financial stability of different age groups is compared and evaluated, to find is there a significant difference in their financial stability. Chi square testing is used as it is a non-parametric testing. The amount of time spend on watching television, by different genders is analyzed. Independence T test is used to compare average score of two different groups. The testing was done to find is there a significant difference in GPA score of males and females.

Part A- Logistic Regression

- **Introduction**

  Logistic regression is a predictive technique used to assess the impact of a set of predictors on a dependent variable (categorical). It helps to test models and predict categorical results with binary, nominal or ordinal values. Binary logistic is used when the Predicted variable has only two outcomes like Yes or No. The plan is to find whether the person is satisfied in life or not given various factors like gender, age, financial stability and all.

- **Research question**

  What components predict the likelihood that person would be overall satisfied in life?

- **Data set Explanation**

  The data was downloaded from Euro-Stat (eurosats, n.d.). The data set contains following variables. The variables 2 to 6 are independent variables and 7 is dependent or predicted variable.
  The satisfaction rating for variables 4,5,6 and 7 (range 1-10, 1-least satisfied and 10-most satisfied)

Serial number	Independent variable	Description
1	Year	The year of survey
2	Age	The age group of the person
3	Gender	Male or Female
4	Financial satisfaction	The rating of how financially satisfied the person is
5	Job satisfaction	The rating of how satisfied the person is with his job
6	Personal relationships	The rating of how satisfied the person is with his personal relationship
7	Overall life satisfaction	Overall life satisfaction rating

- **Principal Component Analysis**

  PCA is used to transform the original number of variances into a smaller set of linear combinations, while retaining all the variance of the variables.

  1. Suitability of data for PCA- Two main issues to consider is sample size and strength of relationship among variables. The table below shows KMO value of 0.718, higher than 0.6 and Bartlett's test sig value of 0.00. Factor analysis is appropriate.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.718
Bartlett's Test of Sphericity	Approx. Chi-Square	1619.224
	df	15
	Sig.	.000

Figure 1

2. Factor Extraction- it is used to find the least number of factors that can be used to properly represent the relationship within variables. The top 3 variables can explain 85% of the variance alone with eigen values above 1.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	2.908	48.458	48.458	2.908	48.458	48.458	2.896
2	1.175	19.582	68.040	1.175	19.582	68.040	1.217
3	1.003	16.711	84.751	1.003	16.711	84.751	1.018
4	.447	7.457	92.209				
5	.271	4.516	96.725				
6	.196	3.275	100.000				

Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

Figure 2

3. Factor rotation and interpretation- factors are rotated either orthogonally or obliquely. This does not affect the underlying solution. The scree plot gives first 2 variables to explain majority of the variance. The elbow rule suggests that top 2 variables are sufficient enough to explain whole data set. But as the variables are less in our data set we will be taking all the variables for prediction.

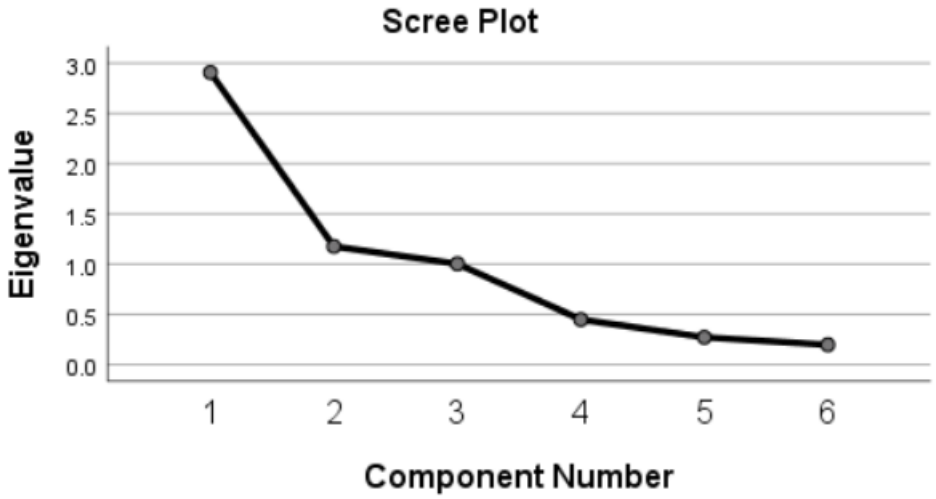


Figure 3

- **Data set analysis**
    1. **Missing values**- There were no missing values. A total of 650 cases were observed.

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	650	100.0
	Missing Cases	0	.0
	Total	650	100.0
Unselected Cases		0	.0
Total		650	100.0

a. If weight is in effect, see classification table for the total number of cases.

2. **Coding of response**- It is used to convert continuous variables to categorical values. For the project three variables were converted. The steps followed to convert in SPSS were TRANSFORM>>>>VISUAL BINNING>>>>Variable to BIN>>>>Make cut-points

Variables	Old value	New value
Age	16-24 years	1
	24-34 years	2
	35-49 years	3
	50-64 years	4
	65-74 years	5
Gender	F	1
	M	2
Overall satisfaction	<=7.4	0
	7.5+	1

Dependent Variable Encoding

Original Value	Internal Value
<= 7.4	0
7.5+	1

Categorical Variables Codings

			Parameter coding			
		Frequency	(1)	(2)	(3)	(4)
AGE_NUM	16-24	135	.000	.000	.000	.000
	25-34	136	1.000	.000	.000	.000
	35-49	136	.000	1.000	.000	.000
	50-64	136	.000	.000	1.000	.000
	65-74	107	.000	.000	.000	1.000
SEX_NUM	F	319	.000			
	M	331	1.000			

Figure 4

- **Assumptions**
  1. **Sample size**-
     The problem with small sample is repetability. The results obtain cannot be repeated with other samples.
     Formula : N> 50+8m (where m= no. of Independent variable). **For this project the sample size is 350. This assumption is satisfied for our sample.**

2. **Multicolinearity**- Collinearity statistics were used to analyse the multicollinearity. The output below shows no multicollinearity present as the Tolerance value is high and VIF value is low.

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	SEX_NUM	.976	1.025
	AGE_NUM	.852	1.174
	financial	.376	2.662
	Job	.350	2.853
	Personal relationships	.480	2.085

a. Dependent Variable: overall life satisfaction

Figure 5

3. **Outliners**- The outliners may severly affect the Goodness of fit of the model. It is important to check for outliners. The outliners are identified by examining the residuals.

Casewise List[b]

Case	Selected Status[a]	Observed JMD1	Predicted	Predicted Group	Temporary Variable		
					Resid	ZResid	SResid
17	S	<**	.965	7	-.965	-5.256	-2.601
62	S	7**	.110	<	.890	2.845	2.125
83	S	7**	.098	<	.902	3.041	2.185
122	S	<**	.997	7	-.997	-16.895	-3.367
135	S	<**	.955	7	-.955	-4.596	-2.504
160	S	<**	.861	7	-.861	-2.488	-2.010
219	S	7**	.056	<	.944	4.112	2.424
451	S	7**	.139	<	.861	2.486	2.027
478	S	7**	.008	<	.992	11.358	3.124
548	S	7**	.025	<	.975	6.226	2.724
618	S	7**	.015	<	.985	8.105	2.906

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2.000 are listed.

Figure 6

- **Estimation of logistic regression model**

Model zero: The model without any independent variable predicted 52.3% times correct result.

$E(y)= e^{-58.124+3.756(finance)+1.92(job)+2089(relationship)} /1+ e^{-58.124+3.756(finance)+1.92(job)+2089(relationship)}$

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted: Overall life satisfaction <= 7.4	Predicted: Overall life satisfaction 7.5+	Percentage Correct
Step 0	Overall life satisfaction	<= 7.4	340	0	100.0
		7.5+	310	0	.0
	Overall Percentage				52.3

a. Constant is included in the model.
b. The cut value is .500

Figure 7

PRESENTING THE RESULTS FROM LOGISTIC REGRESSION

Direct logistic regression was performed to assess the impact various factors have on Overall life satisfaction of a person. The model contained five independent variables (age, gender, financial, job, relationship satisfactions). The full model containing all predictors was statistically significant, Chi square (8, N=650) = 633.216, p < .001, indicating that the model was able to distinguish between people who are satisfied in life with people who are not. The model as a whole explained between 62.2% (Cox and Snell R square) and 83.1% (Nagelkerke R squared) of the variance in life satisfaction, and correctly classified 92.2% of the cases. The table below shows that all variables except gender made a unique statistically significant contribution to the model. The strongest predictor of life satisfaction was financial satisfaction, recording an odds ratio of 42.767. this indicated people who are financially satisfied were over 42.8 times more likely to report overall life satisfaction then those who are not financially satisfied. The odds ratio of age group five that is people over 65 has a value 0.001 which is less then 1 indicating for every year increase in age people are 0.001 more likely to report not satisfied with overall life.

Variables in the Equation

		B	S.E.	Wald	df	Sig.(p value)	odds ratio Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	SEX_NUM(1)	-.335	.321	1.088	1	.297	.715	.381	1.343
	AGE_NUM			73.968	4	.000			
	AGE_NUM(1)	-2.169	.523	17.214	1	.000	.114	.041	.318
	AGE_NUM(2)	-3.122	.571	29.863	1	.000	.044	.014	.135
	AGE_NUM(3)	-4.933	.695	50.434	1	.000	.007	.002	.028
	AGE_NUM(4)	-6.723	.855	61.778	1	.000	.001	.000	.006
	financial	3.756	.408	84.901	1	.000	42.767	19.237	95.074
	Job	1.920	.547	12.320	1	.000	6.821	2.335	19.928
	Personal relationships	2.891	.579	24.923	1	.000	18.010	5.789	56.031
	Constant	-58.124	6.732	74.555	1	.000	.000		

a. Variable(s) entered on step 1: SEX_NUM, AGE_NUM, financial, Job, Personal relationships.

Part B – ANOVA

- **Introduction**

  Analysis of variance is used in comparing mean score of of more then two groups. In this project One-way ANOVA is used. It involves one independent variable with many different levels. It compares the variance between different groups. In this project financial satisfaction score of different age categories are compared.

- **Research question**
  Is there a statistical difference in financial satisfaction score for different age groups?
- **Null hypothesis**- The mean financial score of all the age groups are the same
- **Alternate hypothesis**- the mean financial score of all age groups are different
- **Level of significance**- in this project alpha value of 0.05 is used
- **Data set explanation-**The data was downloaded from Euro-Stat (eurosats, n.d.). The data set contains following variables.
  1. One Categorical Independent variable – AGE with five distinct categories

Variables	Old value	New value
Age	16-24 years	1
	24-34 years	2
	35-49 years	3
	50-64 years	4
	65-74 years	5

  2. One continuous dependent variable – Financial score (range of 1 to 10)

- **Data set Analysis-**
  1. **Missing values**- There are no missing values. The total cases were 680 divided into equal value of 136 each. Hence it is fine if homogeneity of variance assumption is ignored.

Descriptives

financial

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1.0 (16-24)	136	6.349	.8913	.0764	6.197	6.500	3.7	7.8
2.0(25-34)	136	6.197	.7905	.0678	6.063	6.331	3.9	7.4
3.0(35-49)	136	6.074	.9268	.0795	5.917	6.231	3.9	7.5
4.0(50-64)	136	6.039	1.1124	.0954	5.850	6.228	3.6	8.0
5.0(65-74)	136	6.350	1.2297	.1054	6.141	6.559	3.2	8.5
Total	680	6.202	1.0084	.0387	6.126	6.278	3.2	8.5

Figure 8

- **Assumption-**
  1. **Level of measurement**- The dependent variable is a qualitative value at interval level this assumption is met.
  2. **Random sampling**- The scores obtained are from a random sample.
  3. **Independence of observation**- The variables are independent of each other as the ages are not overlapping.
  4. **Normal distribution**- The sample is large hence normality is not checked for the variables.
  5. **Homogeneity of variance**- This assumption is not met by our data set. As the levene's test for equality of variance has a p value of 0.00 which is less then 0.05. we used welch-brown test for significance. P value for the same is 0.027 which is less then 0.05 hence it is significant. Post hoc test for Games-Howell is used as we violated the levene's test.

Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
financial	Based on Mean	10.499	4	675	.000
	Based on Median	10.375	4	675	.000
	Based on Median and with adjusted df	10.375	4	635.019	.000
	Based on trimmed mean	10.560	4	675	.000

Robust Tests of Equality of Means

financial

	Statistic[a]	df1	df2	Sig.
Welch	2.772	4	335.563	.027
Brown-Forsythe	2.918	4	611.892	.021

a. Asymptotically F distributed.

Figure 9

- **Estimation of result**

ANOVA

financial

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	11.736	4	2.934	2.918	.021
Within Groups	678.762	675	1.006		
Total	690.498	679			

Figure 10

The significance value of 0.021 is less than 0.05 hence the test is significant means the sample provided gives evidence **to reject null hypothesis** and say that there is a significant difference in the mean financial score of various groups. To obtain which groups have statistically significant difference **Games-Howell Post hoc test is done**.

Eta= SOS between groups / total SOS

= 11.736/690.498 = 0.027

PRESENTING THE RESULTS FROM ONE WAY ANOVA WITH POST-HOC TESTS

A one way between group analysis of variance was conducted to explore impact of age on financial score, as measured by Europa stat (in range of 1-10). Participants were divided into five groups (Group 1: 16-24 years; Group 2: 25-34 years; Group 3: 35-49 years; Group 4: 50-64 years; Group 5: 65-74 years). There was a statistically significant difference at the $p < 0.05$ level in financial score for the five-age group: $F(4,675) = 2.918$, $p = 0.02$. Despite reaching statistical significance, the actual difference in mean scores between the groups was quite small. The effect size calculated using eta squared was 0.027. Post-hoc comparisons using Games-Howell test indicated that the mean score for Group 1 (M=6.35, SD=0.89) was significantly different from Group 4 (M=6.03, SD=1.11). Overall there was no highly significant difference between groups. The mean

plot below shows the mean comparisons of each group.

Post Hoc Tests

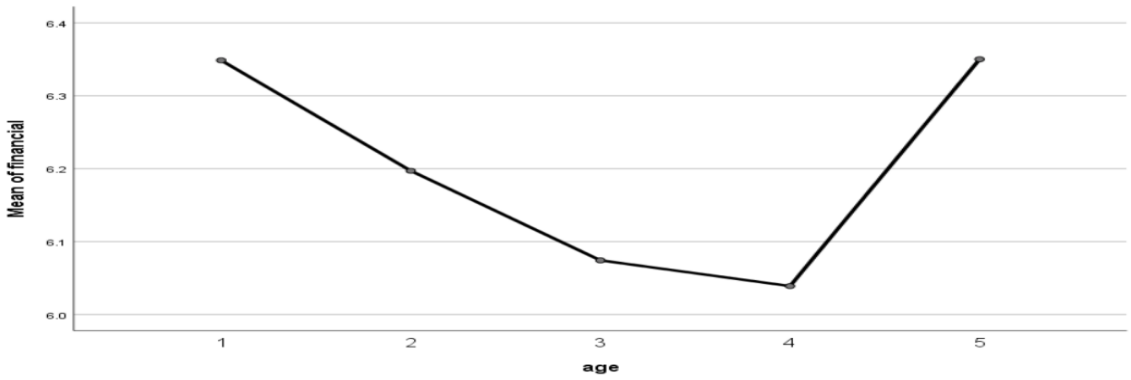
Multiple Comparisons

Dependent Variable: financial

Games-Howell

(I) age	(J) age	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	.1515	.1022	.575	-.129	.432
	3	.2743	.1103	.096	-.029	.577
	4	.3096	.1222	.087	-.026	.645
	5	-.0015	.1302	1.000	-.359	.356
2	1	-.1515	.1022	.575	-.432	.129
	3	.1228	.1045	.765	-.164	.410
	4	.1581	.1170	.660	-.164	.480
	5	-.1529	.1254	.740	-.498	.192
3	1	-.2743	.1103	.096	-.577	.029
	2	-.1228	.1045	.765	-.410	.164
	4	.0353	.1242	.999	-.306	.376
	5	-.2757	.1320	.228	-.639	.087
4	1	-.3096	.1222	.087	-.645	.026
	2	-.1581	.1170	.660	-.480	.164
	3	-.0353	.1242	.999	-.376	.306
	5	-.3110	.1422	.188	-.702	.079
5	1	.0015	.1302	1.000	-.356	.359
	2	.1529	.1254	.740	-.192	.498
	3	.2757	.1320	.228	-.087	.639
	4	.3110	.1422	.188	-.079	.702

Means Plots



Part C- Fundamentals of Statistics

CHI SQUARE TESTING

- **Introduction**
  Chi square testing is used to explore relationship between two categorical variables. It is a non-parametric method of hypothesis testing. In this project it is used to find relationship between gender and sitcoms watched on TV.
- **Research question**
  Is the proportion of females that watch sitcoms the same as the proportion of males?
- **Null hypothesis**- The number of males and females watching sitcoms are same.
- **Alternate hypothesis**- The number of males and females watching sitcoms are different
- **Level of significance**- in this project alpha value of 0.05 is used
- **Data set**- the data set was given on moddle it contains 18 variables and 50 student data.

- **Variables**- in this report the following variables are used
  i) One Categorical variable – GENDER (MALE AND FEMALE)
  ii) Second categorical variable of tvsitcoms- Do they watch sitcom or no (0-NO and 1-YES)
- **Assumptions**-
  i) **Random samples-** Yes, the samples are taken randomly
  ii) **Independent observation**- each student has unique response indicating no repeaters.
  iii) **Unbiased and mutually exclusive**- assumption met
  iv) **Lowest expected frequency** in any cell should be more then or equal to 5

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	11.063[a]	1	.001		
Continuity Correction[b]	9.188	1	.002		
Likelihood Ratio	11.831	1	.001		
Fisher's Exact Test				.001	.001
Linear-by-Linear Association	10.841	1	.001		
N of Valid Cases	50				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 8.64.

b. Computed only for a 2x2 table

- **Estimation of the output:**
  i) **Crosstabulations-** 87.5% females watches sitcoms and 12.5% doesn't. But only 42.3% males' watches sitcoms while 57.7% doesn't. overall 64% of the population watches sitcoms.

gender of student * television shows-sitcoms Crosstabulation

			television shows-sitcoms		
			no	yes	Total
gender of student	males	Count	15	11	26
		% within gender of student	57.7%	42.3%	100.0%
		Adjusted Residual	3.3	-3.3	
	females	Count	3	21	24
		% within gender of student	12.5%	87.5%	100.0%
		Adjusted Residual	-3.3	3.3	
Total		Count	18	32	50
		% within gender of student	36.0%	64.0%	100.0%

  ii) **Chi square test**- In this project continuity correction i.e. yate's correction to compensate for the overestimate of chi square is done as we are using 2*2 table for computation.
  iii) **Effect size-** Phi value of 0.47 indicates near to good association between gender and sitcoms watched.
- **Result**
  The chi square value is significant hence for the given sample we reject the null hypothesis. The alternate hypothesis is accepted. The conclusion is, there is a significant difference in the proportion of females and males watching sitcoms and more proportion of female student watches sitcoms then male.

PRESENTING THE RESULTS FOR CHI SQUARE TESTS

A chi square test for independence (with yate's continuity correction) indicates significant association between students' gender and sitcoms watched, $x^2(1, n=50) = 9.19$, $p = 0.002$, phi = 0.47.

Independent sample t- tests

- **Introduction**
  It is used to compare mean score of 2 different groups of people or conditions.
- **Research question**
  Is there a significant difference in the mean GPA score for males and females?
- **Null hypothesis**- The Mean GPA score for male and female students are same.
- **Alternate hypothesis**- The Mean GPA score for male and female students are different.
- **Level of significance**- in this project alpha value of 0.05 is used
- **Data set-** the data set was given on moddle it contains 18 variables and 50 student data.
- **Variables**- in this report the following variables are used
  (1) One Categorical variable – GENDER (MALE AND FEMALE)
  (2) Second continuous variable of currgpa – The students current GPA score.
- **Assumptions-**
  The basic assumptions are same as Chi square test. Additional assumptions were met and listed below.
  i) **Normal distribution**- The sample size is above 30 and hence this assumption can be ignored.
  ii) **Equality of variance**- Levene's test for equality of variance has a sig value of 0.546 which is > then 0.05 hence this assumption is met for the data set used.

Group Statistics

	gender of student	N	Mean	Std. Deviation	Std. Error Mean
student's current gpa	males	26	3.023	.3983	.0781
	females	24	3.333	.3171	.0647

- **Estimation of the output**:
  Since the sig (2-tailed value) 0.004 is less the 0.05 there is a significant difference in the mean scores of GPAs for male and female students**. For the given sample the null hypothesis is rejected,** and alternate hypothesis is accepted.
- ETA squared calculation
  $t^2 / t^2 + (N1 + N2 – 2)$
  $= (-3.03)^2 / (-3.03)^2 + (26+24-2)$
  = **0.161**
  The above value is greater than 0.14 hence large effect.

PRESENTING THE RESULTS FOR INDEPENDENCE SAMPLE T-TESTS

An independence sample t-test was conducted to compare the GPA score for male and female students. There is a significant difference in scores for males (M = 3.023, SD = 0.3983) and females (M = 3.333, SD = 0.3171); t (48) = -3.03, p = 0.004, two-tailed. The magnitude of the difference in the means (mean difference = -0.3103, 95% CI: -0.5161 to -0.1044) was huge (eta squared = 0.161).

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
student's current gpa	Equal variances assumed	.370	.546	-3.030	48	.004	-.3103	.1024	-.5161	-.1044
	Equal variances not assumed			-3.058	47.023	.004	-.3103	.1015	-.5143	-.1062